数据科学与计算智能：内涵、范式与机遇

程学旗[1]　梅　宏[2,3]　赵　伟[4]　华云生[5]　沈华伟[1]　李国杰[1*]

1 中国科学院计算技术研究所　北京　100190
2 北京大学　北京　100871
3 北京大数据先进技术研究院　北京　100195
4 阿联酋沙迦美国大学　沙迦　999041
5 香港中文大学　香港　999077

摘要　数据科学的发展，将为计算智能的持续发展提供新的可能与机遇；与此同时，计算智能的发展与新型智能范式的兴起，也将为大数据在各行业和各领域的应用提供新的契机。文章阐述了数据科学的内涵，探讨了计算智能的发展与新型智能范式，列举了引领数据科学与计算智能研究的应用方向；进而基于香山科学会议第667次学术讨论会与会专家的讨论，提炼形成数据科学与计算智能领域的七大关键问题，以期使该领域研究得到相关领域研究者与应用者的共同关注，从而把握时代的机遇，推动数据科学与计算智能持续发展。

关键词　数据科学，计算智能，大数据，智能系统，范式

DOI　10.16418/j.issn.1000-3045.20201116005

大数据已成为信息社会的普遍现象，是数字经济的关键资源。以深度学习为代表的大数据驱动的人工智能技术在很多行业和领域获得了成功[1]，这类人工智能本质上源于计算能力，故可将其归为计算智能①。与此同时，大数据是这类人工智能成功的重要因素，这类智能也被称为数据驱动的计算智能，从这个意义上讲，当前数据和智能是一体两面的关系。虽然大数据与计算智能技术在大规模工程化应用方面取得了长足进步，但支撑技术进步的理论基础和技术体系尚处于早期阶段。当前，大数据“红利”效应在逐渐减弱，计算智能技术的单点突破难以为大数据驱动的智能应用提供持续支撑，亟待对数据科学和计算智能的基础问题进行深入思考，重构其理论基石，从而推动技术与工程应用持续进步和跨越式发展。

*通讯作者

修改稿收到日期：2020年11月29日

① 现有人工智能，无论规则驱动或数据驱动，均以计算能力为支撑，本质上是由计算带来的，故称之为计算智能；区分于演化计算领域的“Computational Intelligence”，本文中的“计算智能”英文为“Computing Intelligence”。

本文基于香山科学会议第 667 次学术讨论会与会专家学者的集体智慧，探讨并总结了 4 个方面的问题：① 在数据科学的内涵和外延尚缺乏严谨定义和学界共识的情况下，如何深入认知反映客观世界的数据空间的共性规律？数据科学在本体论和方法论 2 个层面上需要回答的基础问题是什么？② 如何理解、测试并评估现有计算智能的能力边界？人脑、复杂社会系统、自然进化系统等自然智能，往往具备比现有计算智能更加高效的“计算思维”和更加简洁优美的智能推演与决策能力，是否可以借鉴这些自然智能探索新的人工智能范式？③ 在探讨数据科学和计算智能的同时，有哪些值得关注的牵引性应用？新的智能范式对解决复杂的社会问题是否是一个很好的机遇？④ 在未来的发展中，我们该如何把握时代机遇，重点关注哪些关键科学挑战，优先解决哪些关键问题？

1 数据科学的内涵

1.1 基于方法论视角的数据科学内涵

关于数据科学的内涵，一种流行的看法认为数据科学就是图灵奖得主吉姆·格雷（Jim Gray）提出的第四范式（the fourth paradigm）[2]，即在实验观测、理论推演、计算仿真之后的数据驱动的科学研究范式。第四范式的基本思想是把数据看成现实世界的事物、现象和行为在数字空间的映射，认为数据自然蕴含了现实世界的运行规律；进而以数据作为媒介，利用数据驱动及数据分析方法揭示物理世界现象所蕴含的科学规律。这是一种类似方法论视角来定义的数据科学的内涵，即数据驱动科学发现。

第四范式将数据科学从其前的 3 个科学研究范式中分离出来，带来了科学发现和思维方式的革命性改变。借用美国谷歌公司研究部主任皮特·诺维格（Peter Norvig）的话来说，“所有的模型都是错误的，进一步说，没有模型你也可以成功（all models are wrong, and increasingly you can succeed without them）”[3]。海量的数据使得我们可以在不依靠模型和假设的情况下，直接通过对数据进行分析发现过去的科学研究方法发现不了的新模式、新知识甚至新规律[4]。第四范式的一个典型研究案例是关于帕金森病的起因研究[5]。通过对 160 万份病历的大数据分析，研究人员发现帕金森病的起因与人的阑尾有关。这是基于大数据统计帕金森病患病率与切除阑尾的相关性得出的结论。

第四范式通过大数据分析能够发现数据中蕴含的大量相关关系，为科学发现提供了新视野。但是，第四范式本身无法从大量的相关关系中甄别出事物的本质规律。在发现了帕金森病和阑尾的相关性后，有些对第四范式十分执着的学者召集了更大量的帕金森病患者，以彻查他们的基因，调查他们的生活环境和生活习惯，以期从中发现一些共性；然后去找那些也有这些共性但是没有得帕金森病的人，看他们做了什么，有什么共性；如果这种共性存在，可能就是防治帕金森病的解决方案。但是，其结论却不尽人意。可以想象，人体的器官何止一个阑尾，且帕金森病患者的生活习惯何其繁杂，单独靠第四范式的数据驱动方法做漫无边际的相关性分析，不仅要消耗大量的计算资源，也难以真正预测未来的趋势与变化。因此，从方法论来看，第四范式在揭示事物本质规律方面存在固有的局限性，数据科学需要在方法论上突破第四范式。

1.2 基于本体论视角的数据科学内涵

数据科学另外一种值得探讨的内涵是基于“本体论”视角，认为数据是反映自然世界的符号化表示。既然自然世界是客观存在并具备共性科学规律的，那么反映自然世界的数据空间也可能具有独立于各个领域的一般性规律。因而，数据科学应该是“用科学方法来研究数据”，数据科学也应该有类似“信息论”这样的学科基础理论。更具体来看，当我们把世界看成是由物理世界、机器世界和人类社会组成的三元

世界时，新型的“感知、计算、通信、控制”等信息技术使三元世界相互影响和融合，形成了一个平行化（孪生）的复杂数据空间。这样的数据空间，除了映射物理世界，其本身是否具有独特的一般性规律？如何用科学的方法来研究数据的一般性规律，揭示其内在机理？这些是数据科学更基本的问题。例如，数据科学中的一些常数规律（对称性、黄金分割、长尾分布等）和更广意义上的大数据非确定性、数据广义关联、时空演化、数据复杂性等。

1.3 数据科学是方法论和本体论在数据价值实现目标下的统一

数据科学到底应该从哪些视角来定义其独有的内涵与特征？一般认为，作为一门学科的定义，至少应该从其研究对象、方法论和学科目标 3 个维度去界定。数据科学的内涵应该既包括本体论内容和方法论内容，还包括其独特的价值实现目标（图 1）。基于这一认知，可以定义“数据科学是有关数据价值链实现过程的基础理论和方法学，它运用基于分析、建模、计算和学习杂糅的方法，研究从数据到信息、从信息到知识、从知识到决策的转换，并实现对现实世界的认知和操控”[6]。这“三个转换、一个实现”是数据科学的学科目标。而实现这一目标的方法论来自多个学科方法的融合，包括数学（特别是统计学）、计算机科学（特别是人工智能）、社会科学（特别是管理学）等。

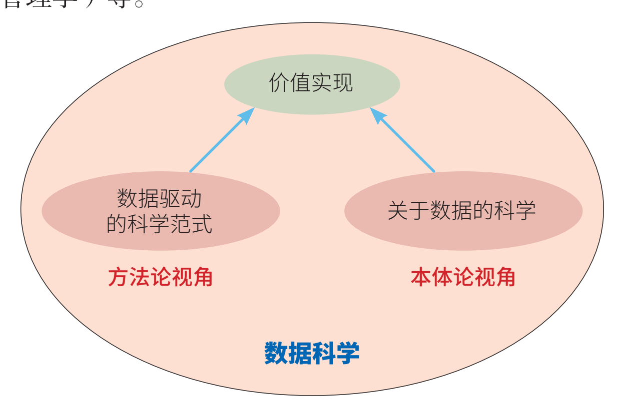


图 1　数据科学的内涵：方法论和本体论在数据价值实现目标下的统一

1.4 数据科学与相关学科的关系

目前，关于数据科学的基本内涵和基础问题还没有像数学、物理学和计算机科学那样成体系、有共识。但是，数据科学的多学科交叉特征及大数据自身的价值特性已经成为共识。我们可以借助相关学科来探讨当前数据科学研究需要关注的基础问题。

（1）*数据科学与统计学*。统计学将数据作为研究对象，致力于收集、描述、分析和解释数据[7]，其为数据科学提供了重要基础和工具。然而，在大数据面前，统计学也面临着诸多问题和挑战。例如：统计假设在复杂大数据分析中难以满足、数据自身及分析结果的真伪难以判定、端到端的大数据推断缺乏基础理论支撑等。统计学针对这些问题目前基本上是束手无策的[6]；而统计学所依赖的一些传统强假设（如独立同分布假设、低维假设等），也都无法适用于目前多源异质的真实数据。因此，数据科学虽然在研究对象上和统计学是相同的，但在研究问题的范畴上却是超越统计学的。譬如：数据科学该如何深入认识数据固有的共性规律？是否能建立一套数据复杂性理论体系？数据规模、数据质量和数据价值有什么定量关系？如何刻画大数据所表现出来的多层面的非确定性特征？

（2）*数据科学与网络科学*。数据科学的发展可以借鉴网络科学的发展历程，以类似的方法寻找研究对象的共性规律[8]。网络科学发现了物理世界中广泛存在的网络所呈现出的共性规律（如幂率分布、小世界现象等），从而促进了其从图论和随机图论中分离出来独立发展，实现了其研究对象从作为数学工具的图到作为物理对象的网络的转变。那么在数据科学中，数据的共性规律是什么？在现实世界中是否有完全不同的两个数据集之间存在某种共性？一方面，一下子找到所有领域的共性规律可能是不现实的，因而可以先从几个关键领域出发，寻找部分领域的共性规律；另一方面，寻找数据的共性规律需要能够问出合

适的基础性问题，类似网络科学中关于度分布、聚集系数、网络直径、网络脆弱性、网络适航性等方面的问题。目前，尚不明确各个领域的数据是否存在统一的规律。因此，数据科学还需要在应用领域进行一定时间的探索，从领域知识中汲取养分，并逐步发现规律、寻找共性。

（3）*数据科学与计算机科学*。数据科学的起源与发展离不开计算机科学，但这两个学科由于研究对象和研究方法的不同，未来也许会平行发展。简单而言，从研究对象的角度来说，计算机科学是关于算法的科学，而数据科学是关于数据的科学。从计算机科学到数据科学，研究手段从传统计算机领域的算法复杂性分析，转变为对数据的复杂性和非确定性等特性进行分析研究。如何对非确定边界的数据，在有限时间空间下进行计算？数据复杂性、模型复杂性与模型性能之间是什么关系？解决某个问题所需要的大数据的量的边界如何确定？是否能发展一套理论，为基于大数据的计算模型提供其能力上、下界的保证？这些都是数据科学独立于计算机科学之外所需要解决的问题。

数据科学目前尚处于发展的早期阶段，其研究方法也应该与传统科学有所区分。数据科学，正处于“无知”到“科学”的中间状态。它目前还没有形成一门完整的学科——信息是不完备的，环境也是非确定的。因此，不能完全按照传统学科来思考和要求数据科学；而应该在这样不完备、非确定的环境下，重新思考和定义数据科学及数据科学亟待关注的基础问题。

2 计算智能的发展与新型智能范式的探索

2.1 计算智能的发展

人工智能（AI）概念在1956年由麦卡锡等学者提出，其发展几经浮沉。基于对智能产生机制的不同理解，人工智能发展至今学派众多，且相互借鉴，形成了一系列代表性成果。无论是早期符号计算（以数理逻辑为基础）、进化计算、支持向量机、贝叶斯网络，还是当前在工业界获得巨大成功的基于多层神经网络的深度学习方法，从模型的本质上来看都是建立在图灵机的基础上[9]，基本都符合邱奇-图灵论题（Church-Turing thesis）[10]，即“任何在算法上可计算的问题同样可由图灵机计算”。换句话说，现有的人工智能模型本质上都是与图灵计算模型等价的，故可归为计算智能。计算智能一般以计算机为中心，以算法理论为基础，充分利用现代计算机的计算特性，给出了解决实际问题的形式化模型和算法。

近 10 多年以来，大数据的使用、算力的提升和深度模型的发展，为计算智能带来了新的契机。大数据、大算力、大模型三者结合，极大地推动了计算智能的工业化应用。例如，计算智能在以围棋为代表的人机对弈、机器翻译、人脸识别、语音识别、人机对话、自动驾驶等应用中均取得了巨大的成功。值得注意的是，大数据在给计算智能带来发展的同时，其复杂性和非确定性也给计算智能带来了非常大的挑战。现有的计算智能在面临大数据环境下的复杂问题和复杂系统时，依然很难给出满意的答案。我们需要探索当前计算智能的能力边界问题，从理论上探寻这类智能所能解决的问题类型和能力边界。譬如，通过建立深度学习和统计力学的关系[11]，回答深度学习的相关基础问题：① 表达能力方面，模型做深为什么是必要的，到底深度为多少层是合理的？② 模型学习方面，崎岖的目标函数如何高效优化？③ 泛化能力方面，如何实现计算智能技术从专用到通用的转变？如何实现模型的跨领域、跨任务、跨模态的泛化？

上述一系列基础问题将进一步成为计算智能未来发展的关键“瓶颈”。其原因是，当前的计算智能是大数据工程化驱动的，其能力的提升主要依赖于数据规模的增加和计算速度的增长。如果缺乏数据科学化理论的支撑，大数据驱动的计算智能难以形成从量变

到质变的提升。那么另一种思路是，我们也许可以考虑发展与当前计算智能不一样的智能范式，以便更加简洁高效地解决更复杂、更普适的现实问题。

2.2 新型智能范式的探索

事实上，自然界中存在大量具备智能的自然系统。这些自然系统比现有人工智能系统具备更加简洁、高效的逻辑推理和自我学习能力，如脑神经系统、社会系统、自然生态系统等。那么，自然系统的智能模型是什么？我们能否借鉴自然系统中的智能行为，将其形式化为可计算的智能范式？实际上，已有4类智能范式在此方面做出了一些初步的探索。

2.2.1 脑启发计算

人类的大脑皮层具有140亿—160亿个神经元，且每个神经元会连接1 000—10 000个其他神经元，借此人类发展出了比其他物种更高级的智慧[12]。脑启发计算（brain-inspired computing）正是借鉴了人脑存储、处理信息的基本原理所发展出来的一种新型计算技术[13]。与传统图灵计算机的计算模式相比，脑启发计算是通过增加空间复杂度来保留计算单元之间的结构相关性，从而构造基于神经形态工程的高速、新型计算架构。脑启发计算的目标是构造一套非“冯·诺依曼”架构、可实时处理复杂非结构化信息、超低功耗的高速新型计算架构。脑启发计算的发展，也许能为数据科学提供新的计算架构和高性能的计算能力，支撑通用人工智能的发展[14]。目前，脑启发计算仍处于起步阶段，我们需要进一步思考如何在不完全了解人脑机制的情况下发展脑启发计算模式，以及如何基于这种脑启发计算为科学研究提供新思路和新范式。

2.2.2 演化智能

学习和演化是生物适应环境的基本方式。现有的计算智能基本都拥有从数据中学习的能力，但对智能模型的演化能力缺乏关注。例如，人脑是经过数百万年的演化逐步形成的。从这个角度来讲，现有的智能模型在依靠人类设计之外，是否也能通过演化过程去自动发现最佳的模型结构[15]？传统的遗传算法是一种基础的演化计算[16]模型；而从演化计算到演化智能，以及实现模型自动演化的智能范式，还有很长的路要走。未来，交互驱动的强化学习、开放环境下的人工智能是值得探索的方向。

2.2.3 复杂系统模拟

自然界存在大量的复杂系统，如人类社会系统、自然生态系统、人体免疫系统等。从控制和计算的角度来看，模型化的复杂系统是“由大量相互作用、相互依赖的单元构成的一个整体系统；一般在没有中央控制情况下，这个整体系统可通过简单的运作规则实现复杂的信息处理，进而产生复杂的集体行为，并能通过学习和进化产生自生长和自适应能力”[17]。是否可以通过模拟复杂系统的组成特点和交互方式来构造新型智能范式？如何通过大量简单智能体之间的交互作用，产生可预期的、具有高度复杂性的群体智能？这样的智能范式也许会从根本上改变传统的单智能体的智能上限。

2.2.4 人机混合智能

随着互联网、物联网及新一代通信技术的发展，万物泛在互联成为现实。未来，大量物理设备、无人系统、人脑，通过泛在网络实现“上线”和“互联”。在这样的环境下，人在回路②的人机混合智能具备了基本的物理条件。目前，人工智能技术所具备的感知、认知能力，基本上是模型与数据结合，并以机器为中心所形成的计算智能，故也称为机器智能。这种机器智能在存储、搜索、感知、确定性问题求解等方面性能表现优越，但在高级认知和复杂问题决策方面与人类智能相差很远。虽然脑启发计算取得了一些进展，但在可预期的未来，机器智能很难完全模仿

② 或称人机互助系统，即人类智能与机器智能形成闭环系统，不断相互作用、相互辅助。

和构造出人类智能或其他自然智能。换一个思路，如果将人的智能引入到机器智能的系统回路中，将充分融合人类智能和机器智能的优势，从而形成更高级的智能水平。在未来较长的一段时间内，这种人机混合智能也许是一些复杂问题求解的有效途径。

那么，在基于机器的计算智能基础上，人作为具备智能的自然系统，如何参与到机器智能的系统回路中是一个关键问题。人机混合智能需要重点解决思维融合或决策融合的问题[18]。具体而言，传统的人机接口往往是单向的；在人机互联情况下，人脑如何参与到机器智能的系统回路当中？如何同时让人理解机器思维和让机器理解人的思维，从而实现思维的无缝互动？目前，一些探索和挖掘思维潜力的工具，如思维导图、思维地图、概念图等，其理论基础与形式化模型并不清晰。一些新型的脑机接口技术进展迅速[19]，但缺乏对人脑在直觉、意识、情感和决策方面的机理认知。也许，从技术上构建有效的人在回路智能通道，是当前人机混合智能亟待解决的关键问题之一（图2）。

2.2.5 小结

上述4类智能范式的研究，在现有图灵等价的计算智能基础上，或多或少地引入了人类智能或自然系统智能的部分机制，从而为未来智能系统的发展注入新的活力。但是迄今为止，这些智能范式在可形式化、可计算、可构造等方面还存在诸多基础性问题挑战。如果这些模式是未来新型智能范式，那么它们是否还是图灵等价的？这些问题值得我们从本源上进行探讨。数据是人类社会、物理世界和机器世界之间的桥梁，同时数据也是人类社会和物理世界的符号化映射。因而，从数据入手是探索和实现上述新型智能范式的基本途径。数据科学基础理论，不仅对当前数据驱动的计算智能起到提质增效的作用，也将为未来新型智能范式研究提供理论支撑。

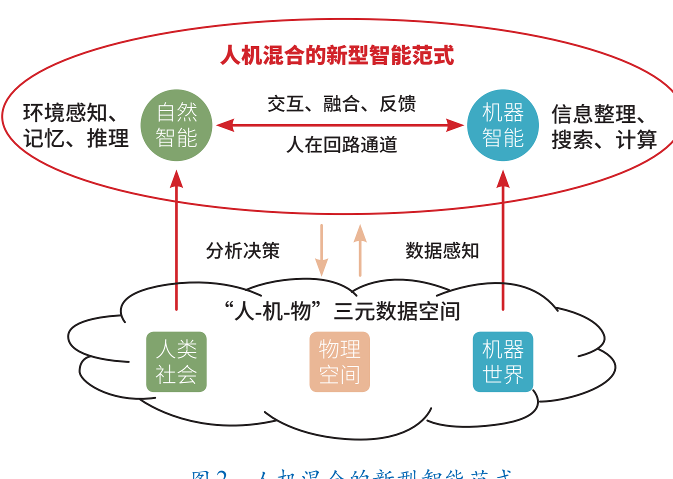


图2　人机混合的新型智能范式

3 引领数据科学与计算智能研究的应用

作为一门实践性强的学科，数据科学的发展离不开实际需求牵引与技术应用驱动。随着感知、计算、通信、控制等技术的发展及综合集成应用，“人-机-物”三元世界高度融合，在线形成了一个网络化的大数据系统，其内部包含了互联网、物联网连接而成的各类数据。这是一个高度复杂、强不确定性、持续动态演化的复杂系统，是“系统的系统”。它既是智慧城市、智能制造、健康医疗等各个领域应用的空间载体，也为国家安全、社会治理、数字经济等领域的科学化、智能化发展提供了重要的数据资源供给。前文已提及，这个现实存在的大数据系统，除了具备高度复杂性、强不确定性等特性，人在回路也是其显著特征。针对这一现实系统的研究与应用，将有可能为数据科学的理论与技术发展带来机遇。针对这一复杂系统的典型场景展开研究，不仅有利于揭示数据的基本规律，也有可能因此而牵引未来新型智能范式的研究。其典型的应用场景有如下4种。

（1）基于非确定数据的社会认知。在社会系统中，我们搜集到的数据通常与真实的情况存在一定的偏差，大量的虚假内容、非确定性内容混杂在这些数据当中[20]。如何能基于这样不完备的、非确定的大数据进行社会认知是一个非常有挑战的问题。社会认知具体包括真假判定、社会心理计算、舆情判定与导向

等。而面向非确定数据的社会认知，其中一大关键在于如何对大量复杂的非确定数据进行假设建模，如何建立复杂社会系统中个人行为与群体社会认知之间的关联。演化智能、复杂系统仿真与模拟也许是解决这一问题的突破口。

（2）基于开放环境的群智决策。互联网极大地方便了信息、知识和智慧的互联互通。在互联网中，已经有许多复杂问题可以通过群智决策的方式加以有效解决，如众包计算、人本计算等。那么，一方面，未来我们该如何设计或改进群智决策中的内部个体交互、融合与反馈方式[21]，以人工构造的群体智能方式进一步提升互联网群智决策的智能上限？另一方面，从计算机的视角来看，该如何利用或者模拟这种人类的群智决策方式，来解决一些复杂的决策问题？考虑到智能系统的演化及复杂系统的仿真与模拟，对单个智能体及智能体之间复杂交互进行建模，也许是未来复杂问题求解的一个可能方向。

（3）人机融合的智慧医疗。智慧医疗是医学、计算机科学、公共卫生学等学科相互交叉的新兴领域。随着信息技术的普及发展，医疗领域产生了大量的数据（如电子病历、PB级基因数据等），也催生了诸多与智慧医疗相关的应用需求。如何根据患者的电子病历及临床影像等数据对疾病诊断提供辅助决策支持？如何根据人类的基因数据，提前进行疾病的预测，为疾病的早期发现、新生儿的先天缺陷预测提供帮助？需要注意的是，智慧医疗需要强大的可靠性，但目前的人工智能还难以替代医生。一种比较好的提高思路是，考虑人（医生）在回路的新型智能范式；通过这样人机混合的方式，使得机器的智能与人的智能相辅相成，使医疗从传统的“个体经验决策”转向“智能辅助决策”的新模式，进而为医疗系统的革新带来新的可能。

（4）重大公共安全问题与社会治理。重大公共安全问题指对社会和公民所需的稳定环境有严重影响的重大问题。公共安全问题涉及多方复杂因素，包括人类社会、自然环境、突发事件等，是典型的人在回路的复杂应用问题，急需应用大数据技术手段进行预测、预警和防控。以新冠肺炎疫情为例，大数据分析技术手段和人机混合智能，为疫情走势预测、传播链排查、谣言传播溯源和意图研判等人在回路的复杂问题提供了有力帮助，支撑疫情精准防控。

4 数据科学与计算智能的关键问题

数据科学的发展，将帮助我们厘清数据科学的理论边界，为计算智能的持续发展提供新的可能与机遇；与此同时，计算智能的发展与新型智能范式的兴起，也将为大数据在各行业和各领域的应用提供新的契机。在本节，我们从数据科学的基本内涵与边界、新型智能范式与智能能力测试、数据评价体系与共享利用3个方面出发，基于香山科学会议第667次学术讨论会与会专家的讨论，提炼形成数据科学与计算智能领域的七大关键问题，以期得到相关领域研究者的共同关注，从而把握时代的机遇，推动数据科学与计算智能的持续发展。

4.1 大数据中的相关关系与因果关系

因果关系指一个变量的发生会导致另一个变量的发生。而相关关系则指一个变量发生变化时，另一个变量也会规律性地发生变化。一般情况下，因果关系往往也是相关关系，而相关关系并不一定是因果关系。大数据的存在，使得人们可以广泛寻求相关关系，Mayer-Schönberger[22]甚至在其书中说道，“大数据时代最大的转变就是放弃对因果关系的渴求，而取而代之关注相关关系”。相关关系确实能在商业和实际应用中带来巨大的成功，但这种成功从科学角度尚需谨慎看待。从科学研究的角度来看，相关关系研究是可以替代因果分析的科学新发展，还是因果分析的补充？从实际应用看，从数据中挖掘出的相关关系能否看作是一种近似因果关系帮助人们进行预测或决

策？对此，不同的学者有不同甚至相反的看法。

建议未来重点研究方向： 相关关系能够逼近因果关系的程度，相关关系和因果关系的边界，是否可以利用反事实推断从相关关系中推断出因果关系，以及如何保证大数据分析的结论可信等问题。

4.2 数据科学的复杂性问题

在计算机科学中，算法的计算复杂性是一个基本问题，包括时间复杂性和空间复杂性。而数据科学除了对计算复杂性的研究外，还需要探索数据自身的复杂性及模型复杂性。数据科学不能一味地靠增加数据量或者模型的参数规模来提升其性能。给定一个具体问题，到底需要多大规模的数据或多复杂的模型才能获得有效解？一个复杂模型判定能力的提升到底有没有尽头或界限？数据规模和模型复杂度之间是什么关系？这些问题在大数据工程化应用中也许可以有经验性的判定，但是在数据科学研究中需要弄清楚其基本内涵和规律。

建议未来重点研究方向： 从数据科学理论出发，给出数据复杂性、模型复杂性和模型性能之间的关系（上下界或渐进理论），为大数据的科学化研究和高效率应用奠定重要基础；当然，要对所有领域给出一个共同的数据科学基础理论，可能比较困难，但可以考虑先从某些重要领域或典型问题出发进行探索。

4.3 有限时空约束下的无限数据计算

在很多场景中，解决问题所需要的数据可能是大量流动的，甚至是无限的——无法确定其边界。例如，真实的自动驾驶技术需要在任意环境、道路上都确保其有效性，理想情况下我们需要通过搜集大量的数据来不断训练自动驾驶模型，促使自动驾驶水平的提升；但问题在于，在实际操作中我们无法在有限时空资源下搜集、处理所有的数据。现有的自动驾驶技术，也基本都是在有限的实验室环境下或者固定的道路上进行学习训练，以期能够实现在任意环境和非确定道路上的自动驾驶。

建议未来重点研究方向： 面向上述边界不确定的数据，到底多大的数据量对问题而言是足够的，以及什么样的数据采样机制才能保证逼近数据整体分布；或者说，该如何在有限时空资源限制下来处理边界不确定的数据。

4.4 强不确定性复杂系统环境下的新型智能范式

大数据空间融合了“人-机-物”三元世界，其交互方式、运行方式极其复杂。复杂系统中跨域高维稀疏的大数据具有很强的时空分布不确定性和价值规律不确定性。在这样一个强不确定性的复杂环境下，能否形成形式化、可计算的新型智能范式？如果存在这样的智能范式，是否还需要依靠大规模数据驱动？现有的脑启发计算、演化智能、复杂系统模拟等主要还是依赖计算机的计算能力，未来还需要进一步探索能够突破计算机计算能力边界的智能范式。人在回路的人机混合智能是一个可能的发展方向，其目标是打通人类智能与机器智能的融合通道，通过有机融合方式实现人机混合智能。

建议未来重点研究方向： 人机混合的智能通道构建及其方式（近几年发展迅速的脑机接口技术、思维融合范式等）；探索这类新型智能范式的主要特征是什么，是否图灵计算等价，是对当前计算智能的改良还是颠覆，以及数据科学在其中发挥什么样的作用等。这些开放性问题研究将为数据科学和计算智能带来新的视野和机会。

4.5 图灵测试以外的通用人工智能测试

图灵测试是早期普遍被接受的人工智能测试准则，主要通过测试者（人）与被测试者（机器）在隔离情况下的问答来测试机器的智能[23]。这是一种非常巧妙的思想实验，但并非工程实验。图灵测试的 3 个开放特点——问题开放、测试者开放、语言开放，导致真正可重复的图灵测试很难实现。而在一般的计算智能设计中，一个重要准则就是需要可重复且有效的评价方式。

*建议未来重点研究方向：*探寻图灵测试之外更加科学有效的通用人工智能测试方法，以及探索以人作为标准答案和参照系之外的可重复且有效的智能评价标准。

4.6 领域无关的数据分类体系与评价指标

数据科学研究中的数据常常来自各个不同的领域，领域之间的数据类型、数据完整性、数据规律等具有非常大的差异性。我们不能只针对某个特定领域的数据来谈论数据科学，而应该对所有领域的数据建立一套共同的话语体系和统一的度量标准。换句话说，需要对不同领域的大数据，进行领域无关的科学分类，构建跨领域、可泛化的数据评价指标和体系。

*建议未来重点研究方向：*可以从数据质量、多样性、复杂性、不确定性或价值密度等多个维度出发，定义数据的统一评价指标。这样的评价指标可以使不同领域的研究者对数据拥有共同话语体系，有利于以数据作为研究对象开展持续的科学化研究。

4.7 可信任的数据共享与流通

大数据是数据科学的研究基础和研究对象，数据科学的发展离不开良性的数据治理和大数据基础环境建设。其中一大挑战问题是可信任的数据共享与流通。数据不同于传统商品，可能会存在无限复制和无限使用的问题，因而造成数据流通价值失效。

*建议未来重点研究方向：*如何用技术手段来确保数据共享和流通的有效与安全，其中数据供给和数据使用是2个关键环节。*① 在数据供给方面，*可以考虑数据的有限供给，通过技术的手段对数据进行限量发行。例如，通过对使用数据的工具增加保护机制，实现数据的有偿服务。也可以利用区块链[24]等技术，保证数据的单方持有。*② 在数据使用方面，*需要考虑数据的有界使用，保证数据的使用不涉及用户隐私等问题。具体来说，可以利用密码学、联邦学习[25]等手段，在保证隐私的前提下加密数据的传输，通过确立数据类型或关系而非获得数据本身作为数据使用的主要方式。数据的共享和流通是数据开放研究的基础，期待未来有更多的人关注数据开放的技术手段研究。

5 未来展望：开启“第五范式”科学研究

在过去十几年间，随着可获得和可使用的大数据持续增长，第四范式作为一种新的科学研究范式，受到科学家越来越多的关注；同时，也暴露出了很多不足。譬如：数据不确定性问题、数据复杂性问题、数据的维数爆炸问题、数据的尺度边界问题等。目前，网络科学、脑科学、社会科学等领域面临的重大问题都是极其复杂且动态变化的难题，采用经典物理一样的简单实验（第一范式）、基于公理和假说的理论推演（第二范式）、基于模型的计算机模拟（第三范式）和数据驱动的相关性分析（第四范式）都无法解决。为此，科学家开始寻求更接近数据和智能本质、更有效认识复杂性和不确定性的新科学研究范式。目前，这类新的科学探索方法论尚未形成定论，大体上看，这类新的科学研究范式是以智能为研究目标的浸入式具身研究，我们暂时称之为“第五范式”③。基于数据科学本体论认识，我们猜测“第五范式”和第四范式一样都会以数据为对象，不同的是“第五范式”更侧重于人、机器及数据之间交互，强调人的决策机制与数据分析的融合，体现了数据和智能的有机结合；“第五范式”强调从本体论的角度看待数据，认为数据本身蕴含自然智能的规律，也是新型智能的载体和产物，期望在数据驱动智能的同时突破现有计算智能的能力边界，借助自然智能构造新型智能范式。

目前，针对“第五范式”的探索刚刚起步，从方法论上还归纳不出它的基本特征；但可以肯定，它的

③ 几年前有学者将“虚拟科学”和“游戏科学”称为科学研究第五范式[26]，与本文提出的“第五范式”的角度有所不同。

一个重要特征是“融合”，既要融合前四种范式，又要融合统计学、网络科学、脑科学等前沿研究中涌现的新方法。第三范式和第四范式都用到计算机：第三范式是“人脑+计算机”，人脑是主角；第四范式是“计算机+人脑”，计算机是主角。第五范式既强调人脑与计算机的“有机融合”，也可能更进一步从社会系统和人脑系统借鉴其中的计算与决策机制，从而更重视人和社会在科学研究回路中的形式化建模与计算融合。

数据科学和计算智能的发展催生“第五范式”；“第五范式”发展离不开对数据科学内涵的丰富和计算智能能力边界的突破。从研究对象看，“第五范式”是科学研究从对物理世界、人类社会的研究拓展到“人-机-物”融合的三元空间；从研究目标上看，“第五范式”不仅仅是传统的科学发现，更是对智能系统的探索和实现；从研究方法上看，“第五范式”强调人在回路的浸入式具身研究。目前，还难以给出“第五范式”的清晰界定，也许再过10—20年，“第五范式”的特征就明朗了，可能逐步成为科学研究的主流范式之一。

致谢 本文的一些观点受到香山科学会议第667次学术讨论会与会者发言的启发，在此对这次会议的所有参加者表示感谢。

参考文献

1 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436-444.

2 Tolle K M, Tansley D S W, Hey A J G. The fourth paradigm: Data-intensive scientific discovery. Proceedings of the IEEE, 2011, 99(8): 1334-1337.

3 Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired, 2008, 16(7): 16-17.

4 李国杰, 程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊, 2012, 27(6): 647-657.

5 Killinger B A, Madaj Z, Sikora J W, et al. The vermiform appendix impacts the risk of developing Parkinson's disease. Science Translational Medicine, 2018, 10: eaar5280.

6 徐宗本, 唐年胜, 程学旗. 数据科学：基本概念、方法论与发展趋势. 北京: 科学出版社, 2020.

7 Moses L E. Think and Explain with Statistics. MA: Addison-Wesley, 1986: 199-203.

8 Barabasi A L. Network Science. Cambridge: Cambridge University Press, 2016.

9 Turing A M. On computable numbers, with an application to the Entschei dungs problem. Proceedings of the London Mathematical Society, 1937, 2(1): 230-265.

10 Church A. A set of postulates for the foundation of logic. The Annals of Mathematics, 1932, 33(2): 346-366.

11 Bahri Y, Kadmon J, Pennington J, et al. Statistical mechanics of deep learning. Annual Review of Condensed Matter Physics, 2020, 11(1): 501-528.

12 尼克. 人工智能简史. 北京: 人民邮电出版社, 2017.

13 Zhang B, Shi L P, Song S. Creating more intelligent robots through brain-inspired computing. Science Robotics, 2016, 354(6318): 1445.

14 Zhang Y H, Qu P, Ji Y, et al. A system hierarchy for brain-inspired computing. Nature, 2020, 586: 378-384.

15 Yao X, Liu Y, Lin G. Evolutionary programming made faster. IEEE Transactions on Evolutionary Computation, 1999, 3(2): 82-102.

16 Back T, Fogel D B, Michalewicz Z. Handbook of Evolutionary Computation. Boca Raton: CRC Press, 1997.

17 Mitchell M. Complexity: A guided tour. Oxford: Oxford University Press, 2009.

18 Sun X, Pei Z M, Zhang C, et al. Design and analysis of a human-machine interaction system for researching

human's dynamic emotion. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, DOI: 10.1109/TSMC.2019.2958094.

19 Wolpaw J R, Birbaumer N, McFarland D J, et al. Brain-computer interfaces for communication and control. Clinical Neurophysiology, 2002, 113(6): 767-791.

20 Güera D, Delp E J. Deepfake video detection using recurrent neural networks// 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Auckland: IEEE, 2018: 1-6.

21 Zhang W, Mei H. A constructive model for collective intelligence. National Science Review, 2020, 7(8): 1273-1277.

22 Mayer-Schönberger V, Cukier K. Big Data: A Revolution That will Transform how We Live, Work, and Think. Boston: Houghton Mifflin Harcourt, 2013.

23 Turing A M. Computing machinery and intelligence. Mind, 1950, 59: 433-460.

24 Underwood S. Blockchain beyond bitcoin. Communications of the ACM, 2016, 59(11): 15-17.

25 McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data// The 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, 2017: 1273-1282.

26 刘益东. 虚拟科学——科学研究的第五范式. 科技创新导报, 2015, 12(29): 7-13.

Data Science and Computing Intelligence: Concept, Paradigm, and Opportunities

CHENG Xueqi[1] MEI Hong[2,3] ZHAO Wei[4] WAH Wan Sang B[5] SHEN Huawei[1] LI Guojie[1*]

(1 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2 Peking University, Beijing 100871, China;
3 Advanced Institute of Big Data, Beijing 100195, China;
4 American University of Sharjah, Sharjah 999041, The United Arab Emirates;
5 The Chinese University of Hong Kong, Hong Kong 999077, China)

Abstract The development of data science is valuable to clarify the theoretical boundary of data science, and provides new possibilities and opportunities for the sustainable development of computing intelligence. Meanwhile, the development of computing intelligence and the emergence of new intelligence paradigms can offer new chance for applications of big data in various industries and fields. This paper discusses the connotation of data science, the development of computing intelligence, the new intelligence paradigm, and lists the key applications leading the development of data science and computing intelligence. Furthermore, based on the discussion during the 667th Xiangshan Science Conference, seven key problems of data science and computing technology are proposed, anticipating to attract attentions of both researchers and applications in related fields, grasping the opportunity of the era, and promoting sustainable development of data science and computing intelligence.

Keywords data science, computing intelligence, big data, intelligent system, paradigm

*Corresponding author



程学旗 中国科学院计算技术研究所副所长、研究员，中国科学院网络数据科学与技术重点实验室主任，大数据分析系统国家工程实验室常务副主任。中国计算机学会大数据专家委员会秘书长，中国中文信息学会信息检索专委会主任。在大数据分析系统、Web 信息检索与数据挖掘等领域发表学术论文 200 余篇，获授权发明专利 60 余项。2014 年获得国家杰出青年科学基金资助。E-mail: cxq@ict.ac.cn

CHENG Xueqi Full Professor and Deputy Director of the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). He is the Director of CAS Key Laboratory of Network Data Science and Technology. His research interests include big data analytics, Web search and data mining. He has published more than 200 papers, and has more than 60 authorized patents. He is the Secretary General of Big Data Society, China Computer Federation. He is also the President of the Society of Information Retrieval, Chinese Information Processing Society of China. He was funded by the National Science Fund for Distinguished Young Scholars of National Natural Science Foundation of China. E-mail: cxq@ict.ac.cn



李国杰 中国工程院院士、发展中国家科学院院士。中国科学院计算技术研究所原所长、研究员，中国科学院科技战略咨询研究院科技智库特聘研究员。1943 年出生于湖南，1985 年在美国 Purdue 大学获得博士学位。主要从事并行算法、高性能计算机、互联网、人工智能等领域的研究，发表学术论文 150 余篇，出版《创新求索录》个人文集。主持研制“曙光-1000”等计算机，获国家科技进步奖一等奖等奖励。E-mail: lig@ict.ac.cn

LI Guojie Born in 1943, received his Ph.D. in 1985 at Purdue University, USA. He was the director of the Institute of Computing Technology, Chinese Academy of Sciences (CAS), and now is a professor of this institute and a specially-appointed research fellow of the Science and Technology Think Tanks in Institutes of Science and Development, CAS. He mainly engages in researches on parallel algorithm, high performance computer, internet, and artificial intelligence. He has published more than 150 academic papers, directed a series of projects such as building Dawning-1000 computer, and won the First Prize of National Science and Technology Progress Award. He is a member of Chinese Academy of Engineering, as well as fellow of The World Academy of Sciences for the advancement of science in developing countries (TWAS).
E-mail: lig@ict.ac.cn

■责任编辑：文彦杰